Mathematics, Computing and Technology: Level 2
M248 Analysing data

The Open University

M248
DVD Guide

Contents

SUP 54538 6
1.4

Introduction

These notes accompany the collection of seven TV programmes that are provided on DVDs as part of M248 *Analysing data.* The programmes are all documentaries. They explore the extremely wide range of important and interesting applications that statistics has in the modern world. The application areas range from public health and medicine, to ecology, the steel industry and on to major areas of public interest such as the interpretation of performance indicators and the public perception of risk. There is also one programme (TV7) of a historical nature, exploring the life, achievements and legacy of Sir Ronald Aylmer Fisher, the great twentieth century statistician and geneticist; this also shows how important Fisher's methods and ideas remain in the modern world.

The benefits you will get from watching these programmes include

- ◇ a better understanding of the role of statistics and quantitative thinking in public life and the media;
- ◇ a good idea of how statisticians cooperate with researchers in other disciplines to help answer real life questions;
- ◇ the chance to see many of the techniques that are developed in the written materials of this course applied in practice;
- ◇ the opportunity to see for yourself the diverse and important roles of statistics — and of statisticians — in the world today.

Tracks 4–7 have been made recently. The first three of these programmes form a short series entitled *Vital Statistics.* The other three programmes (Tracks 1–3) were made in the mid-1990s for the predecessor of this course, M246, but remain of equal benefit and relevance to the current course.

It is hoped that as well as adding to your understanding of statistics, these programmes will be both enjoyable and stimulating in showing the contribution of statistics, in particular through analysing data, to solving important problems.

Track 1 Ecological predictions

Figure 1 Monitoring a randomly walking newt

This course is about data and models for the data. But what if you cannot realistically get all the data you would like? And what if the models you would like to fit lead to mathematics that is just too difficult to handle? In both cases, a possible solution is to turn to the computer and to *simulate* the situation of interest.

Ecological systems are usually very complicated. Models of such systems can be built, often utilizing simple modelling components. But the large scale of the systems allied with a mass of complicated interactions makes for model intractability. Obtaining data to answer your question directly is typically impossible too: ecological time scales are usually too large, and experimentation may well be too destructive. So, as presenter David Hand found out when he visited Little Wittenham Nature Reserve, near Oxford, simulation is a tool much needed and used by ecologists.

David describes three such applications at Little Wittenham. They concern newts, grasses and trees, respectively. The first application concerns the Great Crested Newt, a protected species. Newts are amphibious, breeding in ponds, but also spending much time on land, feeding and then hibernating. On awakening from hibernation, newts walk in search of a pond — perhaps the old pond, perhaps a new one — in which to restart the breeding process. Now, as ecologist Paul Frankling shows, newts tend to walk in apparently random directions, with a resulting overall speed averaging something like 2 metres per hour. Notice the random element here. It is *statistical* simulation because the model being simulated involves a random element. (It can also be valuable for mathematical modelling exercises to use the computer to simulate complicated *deterministic* models.) This means that for any one simulation, the precise directions and distances travelled by a simulated newt arise from random numbers provided by the computer. In fact, the angles at which the newts set off are simulated from the continuous uniform distribution.

The more ponds within the newts' range, the more likely a newt will 'move' to a pond different from the original one, and the more the species will be able to spread. The question this simulation attempts to answer, therefore, concerns the distance there needs to be between ponds for the newts to have a good chance of proliferating.

The second application of simulation addresses competition between species of grasses. Over the years, intensive farming practices have had a considerable effect on native wildlife communities. More recently, however, farming policies have changed, with traditional crops giving way to other crops or simply to no production. Fiona Woolmer and Cathy Lines are both interested in grasses, with a view to studying and predicting the restoration of 'impoverished' grassland. In particular, they would like to know which, if any, grass species will come to dominate when several species are in competition. Cathy's research provides numbers for the 'invasion rate' of one species with respect to another for each of five main grasses of interest. These, like the speed of 2 metres per hour provided by the ecologist's work in the newt context, become values for parameters of a suitable ecological model. But the model is too complicated to pursue other than by simulation. And the simulation yields striking and informative visual results.

For the third application we turn to forestry. If trees are planted too close, too much competition for space and light will result in few trees attaining sufficient size for the timber industry. If they are planted too far apart, on the other hand, the trees may get too bushy, and the land is used inefficiently. Using the specialist knowledge of contributor Gabriel Hemery, a simulation can be set up to try to pick an optimal spacing for the ash trees he wants to plant at Little Wittenham. Here, random numbers drive the starting heights of the young trees and are combined with an established deterministic model of tree growth.

If simulations are so useful, you might ask, why do we not simulate everything? Well, simulations have to be rerun for each choice of model parameters, and repeated many times to gain representative behaviour accounting for random variation. Making sense of results can, therefore, be difficult. A theoretical analysis, when available, often offers more insight into the underlying processes and more readily generalizable results. From the data viewpoint, models imply assumptions and simplifications that may not be tenable in the real world: real data, on the other hand, directly address the actual problem of interest rather than its idealization. But for all that, statistical simulation remains a very valuable tool, often complementary to other approaches, and sometimes the only way forward.

Presenter: Professor David Hand.

Other contributors: Paul Frankling, Fiona Woolmer, Dr Cathy Lines and Gabriel Hemery.

Narrator: Veronica Hyks.

Producer: Andrew Law.

Track 2 Clinical trials

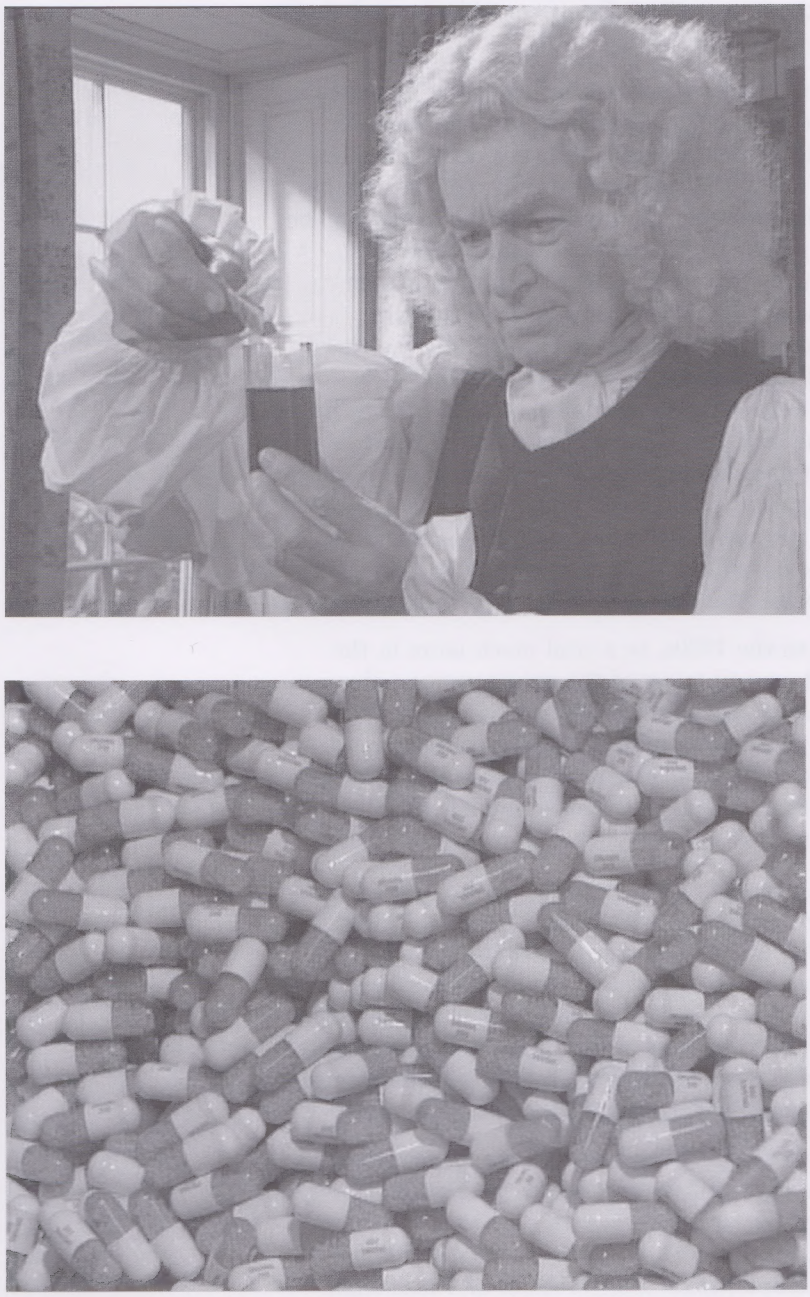

Figure 2 Mid-eighteenth and twenty-first century medicine

To test whether or not a new drug works in practice, a study known as a clinical trial must be carried out. However, as with many other things, the results of applying the drug to a range of patients will not be clear-cut — neither 100% of patients will, in general, be subjects of a miracle cure nor, one very much hopes, will everybody suddenly get more ill or die. As a consequence, there is a major input of statistical science into such trials. Nowadays many such trials are being performed at any time, each and every one subject to serious scrutiny and rigorous analysis. The job title 'medical statistician' accounts for an important and sizeable proportion of all statistics professionals.

To explain some of the main issues in the statistical analysis of clinical trials, in this programme David Hand leads us through three historical trials that represent major stepping stones in clinical trial development. In 1763, the Reverend Edward Stone, vicar of Chipping Norton in Oxfordshire, sent a paper to the Royal Society describing the results of his previous five years' work (one of the earliest

clinical trial reports, although Stone would not have used such terminology). Stone was concerned for his parishioners and, in particular, was seeking out natural substances for potential use as cures for the many agues and fevers that they had. Of particular interest was willow bark. Stone gave ground willow bark, a powder served in a drink, to 50 patients over those five years and reported that all had reacted favourably.

Does this prove anything? Well, Stone had been aware of some of the ingredients of the modern clinical trial, but by no means all. It seems that (i) he was well aware that people were different, and just one patient's recovery does not imply everyone's recovery in the same way, hence the sample of size 50; (ii) he took into account the possible effects of the drink he used (perhaps it was the drink not the powder that was responsible for any perceived improvement) and varied the drinks used to check this out; and (iii) he was aware of the possibility of side effects and reported that there were none. But some things were less well accounted for: (i) perhaps all those patients would have recovered anyway (it is not clear how serious the agues were), there being no kind of 'control' patients — those not taking the drug — to make comparisons with; (ii) try as he did to be fair, it is possible that Stone was subconsciously affected by what he wanted to observe; and (iii) perhaps taking an unpleasant medicine had a positive (psychological) effect on the patients rather than there being a true effect of the particular treatment. As it happens, the active ingredient that Stone had stumbled across *was* effective, as is now well known, and continues in very widespread use to this day: find out what this was by watching the programme!

From 1763 the programme moves on to the 1930s, to a trial much more in the modern mode, with much attention drawn to many of the considerations above, but with one remaining flaw which could well have invalidated the project's results. The issue here was the effect of drinking milk on children's physical development. The trial — perhaps not quite what everyone would call a 'clinical' trial since the 'drug' was milk, but clearly a trial in the same mould, with consequences for health — was performed in schools in Lanarkshire, Scotland.

The difference from Stone's trial was that the effects of interest were likely to be fairly subtle and long term, as in many modern medical investigations. (Stone's results stood up partly because *all* his patients recovered: perhaps not a miracle cure, but nonetheless a persistent effect.) Such subtle effects were detected by taking before-and-after weight measurements on the children over a suitable time period and by averaging results over large numbers of children. The comparison problem that Stone did not address was here attended to by splitting the children into two groups: one group received the treatment (milk), the other did not.

The stage was set for analysis by a two-sample t-test, an important technique discussed in *M248*, which is explained here and brought to life by a graphics sequence. As is usual in hypothesis testing, a 'null hypothesis' was set up. Here, it was assumed that the treatment had no effect. Then the expected distribution of the difference between the means of samples from the two (milk and non-milk) populations was calculated. Finally, the observed difference in the trial was compared with this sampling distribution to see whether the observed difference was typical or unusual: if the former, it would be said that there was no evidence to suggest that anything other than random variation was needed to account for the observed difference, while if the latter there would be evidence against the null hypothesis. For the Lanarkshire milk experiment, the observed difference in children's weight gains would occur less than 1% of the time if the null hypothesis were true: so, we have strong evidence that milk was effective.

So, how were the two groups chosen? Although the basic idea was to assign children to groups randomly, teachers were given some freedom to change things to 'balance' groups. It seemed that in some cases teachers with a belief in the power of milk tended to select and give the milk to the less well-nourished children. This means that the groups differ to start with — so casting doubt on the results.

The third and final clinical trial David describes is a little newer, but very important, and suffers no obvious drawbacks. It involves streptomycin, the antibiotic treatment for tuberculosis (TB). In the 1940s, the initial clinical trial of

streptomycin compared a group of patients receiving bedrest only (the standard 'treatment' previously) with a group receiving streptomycin as well. A shortage of the drug meant that the trial was limited to only a few hundred patients. But, as well as the basic safeguards, the investigators took care to limit the variability of their results by concentrating on one age range of patients and on certain restricted forms of TB. Just 7% of those treated with streptomycin died (in a certain time interval) compared with 27% of the others, and the significance probability was under 1%: there was evidence that the new treatment worked.

David ends with a summary of the basic statistical requirements of modern clinical trials. These include (i) sufficient sample sizes, (ii) proper randomization, (iii) the eradication of psychological patient effects by (a) the giving of inactive substances called placebos, (b) the 'blinding' of the patient to which treatment he or she receives, and (c) the eradication of potential investigator biases by also 'blinding' the administering doctors to which treatment is which. And, as always, appropriate statistical analysis of the data!

Presenter: Professor David Hand.

Other contributor: William Russell as Reverend Edward Stone.

Narrator: Veronica Hyks.

Producer: Andrew Law.

Track 3 Regressing to quality

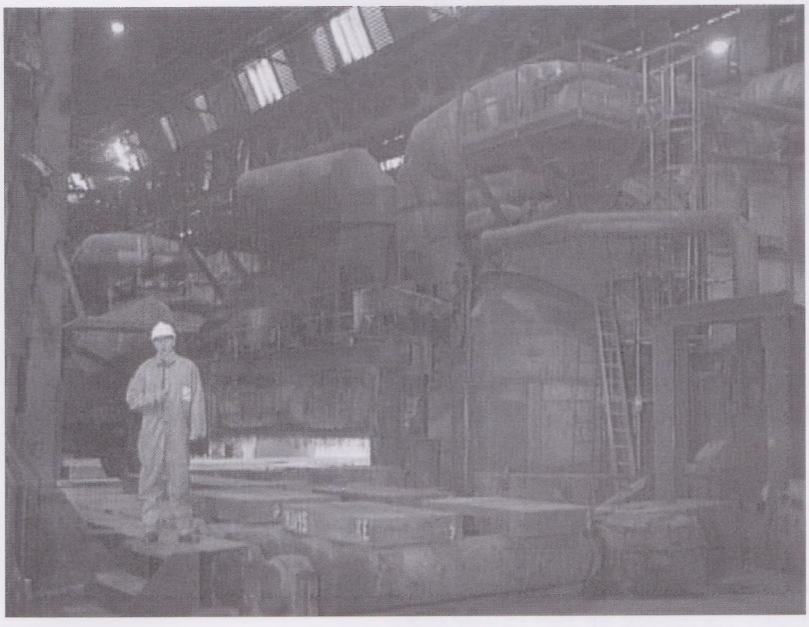

Figure 3 David Hand and a giant furnace

Manufacturing industry and statistics, you might think, make strange bedfellows. However, from the second half of the twentieth century into the twenty-first, nothing could be further from the truth. Quality is an industrial buzzword and assessing, improving and maintaining quality of processes and products is an area with immense statistical input. For many years, topics such as 'statistical quality control' have been used in a monitoring capacity to check that products meet specifications. But statistics nowadays also plays a part in a whole new ethos of industrial management which takes account of quality at every step of the manufacturing process, so-called 'total quality management'. Indeed, it is often

claimed that the success of Japanese industry since the Second World War is founded on approaches like this, approaches with a strong statistical influence.

Statistical science is certainly used to good effect by British Steel, as this programme shows. David Hand follows the story of the production of oil and gas pipes from their beginnings as iron ore to their finished state for installation in the North Sea. As British Steel expert Chris Bennett tells us, these are no ordinary pipes. They need to be strong enough to resist damaging impacts and to survive massive changes in pressure, and they need to be clean and resistant to impurities in the materials they transport. The requirement is thus for the pipes to be made of high-quality, yet cost-effective, steel and also for high-quality manufacturing of the pipes from the steel.

David tells the story stopping at three distinct points where statistics has played an important part, although these are by no means the only places in the process where statistics plays a role. All three statistical analyses involve the very important statistical technique of regression analysis.

The first stopping point in the programme concerns the control of impurities in making the steel. For many purposes, the maximum allowable sulphur content is just 0.001%. But there is rather more sulphur in the basic iron than this. Sulphur content is reduced by injecting magnesium which reacts with the sulphur to create a slag of magnesium sulphide which can be removed. Given the sulphur content of a batch of iron ore, what is the minimum amount of magnesium needed to reduce the steel sulphur content to the required level? It is important to be able to achieve the minimum because magnesium is very expensive. Sulphur content (untreated) is thus the explanatory variable in a simple regression set-up with the amount of magnesium needed being the corresponding response variable. The desire is to obtain a regression relationship to predict the future magnesium requirement from any future iron sulphur content. The data come from previous magnesium determinations in which a little magnesium was added at a time (a slow and inefficient process). The resulting linear predictor is not an exact relationship between sulphur and magnesium in a deterministic way but a good estimated predictive relationship with a certain amount of inevitable random error. Nonetheless, it turns out to yield an extremely useful and simple way for British Steel to determine how much magnesium to use.

The programme refers to *predictor* variables instead of explanatory variables. Both terms are in common use in statistics, and (in this context) mean the same thing.

Later on in the production process, slabs of steel need to be heated to a high temperature in a furnace, ready to be rolled into an appropriate thickness to be made into pipes. Here, it is imperative that the slabs are heated to the same temperature throughout. This calls for giant fuel-greedy furnaces, which inevitably waste a certain amount of energy. From time to time, the furnaces are refurbished in an attempt to save fuel. But how can British Steel check whether the refurbishment has worked? Different amounts of throughput occur at different times, so they cannot just look at their fuel bills. The answer, again, is to use regression. Both before refurbishment and after, measurements are made of throughput (the explanatory variable) and of energy consumption (the response). For each set of 'before' and 'after' measurements, linear regression seems appropriate to fit. The answer is a clear saving at each and every level of throughput (at something around a level of 6%).

Regression on one explanatory variable, separate regressions on one explanatory variable for each of two groups, and finally regression on two explanatory variables: David finds himself moving from Scunthorpe to Hartlepool to observe how the pipes are 'sewn up' by welding. Again, the quality of the weld has to be extremely high. One particular aspect of interest is the carbon content of the weld: this should be at a certain level. British Steel can avoid much costly and time-consuming inspection of welds by employing a regression model to predict weld carbon from two other carbon measurements, that in the steel plate and that in the wire used to make the weld. Least squares is used to fit a plane which predicts weld carbon for any combination of plate carbon and wire carbon. This is used to determine which type of wire to use in the welding process, given the amount of carbon in the plate, and the desired resulting weld carbon level.

Watch out throughout for striking graphical illustrations of how and why lines and planes may be fitted by least squares.

Presenter: Professor David Hand.

Other contributors: Chris Bennett, Nick Woodford and Jack Dainty.

Narrator: Veronica Hyks.

Producer: Andrew Law.

Track 4 Something in the air?

Figure 4 The Broad Street pump: symbol of public health epidemiology

Do mobile phones harm the brain? Is genetically modified food bad for you? What impact does a sedentary lifestyle or environmental pollution have on your health? What are the risks and benefits of vaccination? Nearly every week some new health-related question receives prominent media coverage. But who are the people studying these issues, and how do they go about it? In this programme, presenter Jancis Robinson illustrates the part played by statistics in investigating health matters by looking at three very different topics: asthma, cholera and lung cancer.

The programme opens with the Ramanoop family. Parents Wendy and Alan are talking about their daughters Victoria and Kimberley, who suffer from asthma. They are not alone: asthma rates have been rising for many years in Britain and other countries. However, what is causing this increase is not well understood. Various theories have been proposed, for example, that the rise might be linked to increased air pollution, to factors that induce allergies, or to changes in our diet.

To examine the issue it is necessary to consider not just individual cases like the Ramanoops, but to look more widely at the patterns of asthma in the population as a whole, in the hope that such patterns or trends might throw light on possible causes. The study of diseases in populations is called epidemiology. However, revealing and interpreting patterns in data is the subject of statistics. Hence statistics plays a central role in epidemiology.

Having set the scene with the Ramanoops, presenter Jancis Robinson looks back to the origins of epidemiology. She goes to meet epidemiologist Rosalind Stanwell-Smith in the John Snow pub in Soho, London. John Snow was a physician who lived in London in the mid-nineteenth century. At this time, regular cholera epidemics occurred with devastating mortality, but their cause — as with asthma today — was not fully understood. The prevalent theory was that the disease was caused by 'miasma', or bad air, that accumulated in low-lying places. John Snow disagreed, believing instead that cholera was related in some way to consumption of contaminated water (this was long before germs were discovered).

Snow set out to test his theory using statistical methods, by collecting data and looking for patterns. At that time London's water was in part supplied by private companies, two of which, the Southwark and Vauxhall Company and the Lambeth Company, drew their water from polluted parts of the river Thames. In 1853, the Lambeth Company moved its water intake upstream to a less polluted part of the river. The following year, a cholera epidemic struck. Snow reasoned that if cholera was related to water contamination, there should be fewer cholera cases, in proportion, among customers of the Lambeth Company compared to those of the Southwark and Vauxhall Company. Snow went from door to door, collecting information on cholera deaths and water sources. He found that there were indeed far more deaths among customers of the Southwark and Vauxhall Company: the actual figures he collected (not given in the programme) were 315 deaths per 10 000 houses compared to 38 for the Lambeth Company.

Snow also plotted the location of the cholera deaths on a map. He found that many deaths occurred in the Golden Square area of Soho, and appeared to be clustered around the pump in Broad Street where the John Snow pub is now situated. He deduced that the water from this pump was also contaminated, and removed the handle from the pump to stop anyone else from using it. This act has gone down in history as a defining moment for public health epidemiology, a key aspect of which is to put the knowledge gained from statistical and other data to direct use through public health interventions.

Returning to the asthma story, Jancis Robinson visits Paul Aylin who is studying patterns of asthma in relation to air pollution. Paul describes the types of statistical methods he is using, some of which are discussed in the course. One approach is to look at trends in asthma rates over time to see if they peak during periods of high air pollution. Another is to compare rates of asthma in different geographical areas, to see if more polluted areas have higher asthma rates. However, he warns against concluding too hastily that a pattern, or association, necessarily implies a causal link. For example, apparent associations might be due to chance, or bias. Alternatively, they might be due to confounding effects: for instance, polluted and unpolluted areas may differ in other respects, and it may be one such other factor, not the pollution, that is causing the asthma rates to differ. It is only once these other factors are ruled out that it is possible to conclude that an association may be causal. The important distinction between association and causation is one with which you will become familiar during the course.

The need to keep an open mind is reinforced by Anthony Seaton at Aberdeen University, who is studying the impact of diet on asthma. He has noticed that our consumption of fresh vegetables has declined dramatically over recent decades, and has suggested that this, particularly changes in our intake of Vitamin E, might be linked to asthma. However, he is careful to emphasize the need to test his hypothesis repeatedly in different studies. Testing hypotheses is an important aspect of statistics.

Such a painstaking, thorough, statistical approach can yield huge benefits, as illustrated by the issue of smoking and lung cancer. Jancis Robinson meets Sir Richard Doll, who in the 1950s first identified the link between smoking and lung cancer. At the time, a big increase in the rates of lung cancer had been noticed, but it was not clear what was causing this rise. It was thought that, perhaps, air pollution was to blame — London, in particular, was notorious for its

smogs — or that the coal tar used to tarmac the roads was the cause. Together with statistician Austin Bradford-Hill, Richard Doll undertook several statistical investigations which showed that the cause was, in fact, smoking. This conclusion has since been confirmed by many further studies, though it was not until the 1970s that smoking began to decline.

Statistical investigations are central to uncovering the causes of diseases in populations. To underline the point, the film shows Patricia Kane, a research nurse specializing in respiratory disease, collecting data from the Ramanoop children, as part of a statistical study into the causes of asthma. A full understanding of the causes of asthma, like the link between water and cholera uncovered by Snow, or that between smoking and lung cancer revealed by Doll, still eludes us. But one thing is certain: statistical methods will play a key role in solving the riddle of asthma as well.

Presenter: Jancis Robinson.

Academic consultant: Dr Paddy Farrington.

Other contributors: The Ramanoop family, Dr Rosalind Stanwell-Smith, Dr Paul Aylin, Professor Anthony Seaton, Patricia Kane and Professor Sir Richard Doll.

Producer: Alison Priestley.

Track 5 Risk

Figure 5 Will she jump? Watch the programme to find out ...

Sarah Miller is about to make her first parachute jump. She has thought through the implications and weighed up the risks. Crouching by the open hatch of the plane, she can see the airfield far below, where her son is waiting for her. Now she has to decide, finally, whether or not to take the jump.

Every day of our lives we have to make decisions. Many are trivial, some momentous. All are taken in conditions of some uncertainty. How do we go about making decisions? What influences us? And how can statistics help us make better choices?

Peter Bernstein explains that it is best to think of risk as a combination of probability and consequence: probability that the event will happen and consequence of it happening. When making decisions, we are confronted with the inevitable fact that we do not know what is going to happen: we must therefore weigh up our decision in the light of its likely consequences. He gives the example of crossing a busy road: the probability that he will be hit by a car is small, but the consequence of being hit could be very serious. So he waits for the lights to change. The key point is that consequences dominate in the decisions we make.

David Hand describes how, when we make decisions, we subjectively combine probabilities and consequences, which might include benefits as well as costs. But how do we balance all the different factors, and what influences our decisions? Sarah Miller's decision to do a parachute jump is a personal choice, made despite the fact that she is afraid of heights. Sarah does not fully understand her own motivation, and the film follows her as she comes to terms with the implications of her decision. She learns about the dangers of parachute jumping, which can result in injury, or even death. She practices what to do if she is unlucky enough to be the 1 in 750 for whom the parachute does not open normally. She discusses her decision with other parents as she collects her son Thomas from school.

The consequences and impact of our decisions vary according to our circumstances. David Hand describes why he gave up cycling in London after he became a father: though the probability of an accident had not changed much, the consequences had become more serious. In the language of risk, becoming a parent had made him more risk averse. Sarah Miller also reflects that she would need to think of the consequences for her son were she to consider taking up parachute jumping more regularly. For others, however, and in particular for the women who are taking them up in increasing numbers, dangerous sports like parachute jumping bring unique exhilaration and excitement.

Many decisions involve balancing different risks. We meet Sousan Azimrayat who is taking her daughter Gina to be immunized against polio, a devastating disease that can cripple children for life. Thanks to vaccination, polio is now virtually eliminated in Britain, but in about one in a million instances a child can catch polio from the vaccine. Thus, vaccination carries a risk, albeit a small one. David Elliman, from St George's Hospital, London, explains that the choice parents face is not between taking or not taking the risk to have their child vaccinated, but involves balancing different risks.

Sousan and her partner, Peter, are also well aware of another dimension to the vaccination issue, namely the benefit of vaccination to a society as a whole, and to other children. Their first child, Thomas, was born with a serious heart condition, and could not receive the vaccine. However, as explained by David Elliman, Thomas was protected by the fact that a very high proportion of the population is vaccinated, thus reducing the probability of catching the disease for the few remaining unvaccinated.

Statisticians have developed methods to formalize the process of decision-making. One approach is to quantify the benefit or desirability of each outcome mathematically; this is known as the utility of the outcome. These utility values, which presenter Jancis Robinson briefly alludes to in the film, are then combined with the probabilities of the various outcomes and the costs involved in each decision. The aim of the analysis is to identify the optimal course of action. While we do not usually employ such formal methods in everyday life, they are commonly used in some fields of activity. In particular, the insurance industry typically uses statistical methods of assessing risks and costs, which in their case can be quantified in terms of financial costs and benefits.

Peter Bernstein explains that to take out an insurance policy is to transfer some of the consequences of a risk to others, which of course comes at a price. He goes on to compare the operation of an insurance company to that of a casino. When you take out a life insurance policy, for instance, you are betting against the insurance company that you will die before your time. If you die early, you win your bet and the insurance policy pays out to your family. Clearly, insurance

companies carefully weigh up the probabilities and the risks so that, on average, they make a profit.

Sousan has experienced directly the impact of such calculations. In October 2000, Sousan's electronics company in Lewes, East Sussex, was flooded when the river Ouse burst its banks. Now Sousan is finding it impossible to obtain flood insurance for her premises, as the insurance companies have reassessed the risk of flooding and the likely cost to them.

Jancis Robinson goes to meet Peter Midgeley, the Environment Agency officer responsible for the flood defences in East Sussex. He describes the October 2000 floods as a once in 200 years occurrence: the floods were the most serious since records began nearly two hundred years ago. While the 1/200 annual probability is indeed based on historical records, Sousan disputes its validity today, a view apparently shared by the insurance companies who are no longer prepared to insure against flooding.

As circumstances change, the probabilities upon which our decisions are based may also need to be adjusted. Peter Bernstein briefly describes a statistical approach to updating probabilities, based on an original idea of Thomas Bayes, an eighteenth century clergyman. This is a method for changing the probabilities by combining them with new data, thus allowing risk assessments to be updated as new information becomes available. Thus, for example, the risks of flooding can be reassessed to take into account changed circumstances such as global warming, as well as historical data.

Back at the airstrip, the weather has settled. Has Sarah assessed her utility values and updated her probabilities? To the sounds of an ominous sound-track, she takes to the air. But will she finally make that jump?

Presenter: Jancis Robinson.

Academic consultant: Dr Paddy Farrington.

Other contributors: Sarah Miller, Sousan and Peter Azimrayat, Professor Peter Bernstein, Professor David Hand, Dr David Elliman and Peter Midgeley.

Producer: Alison Priestley.

Track 6 Hitting targets

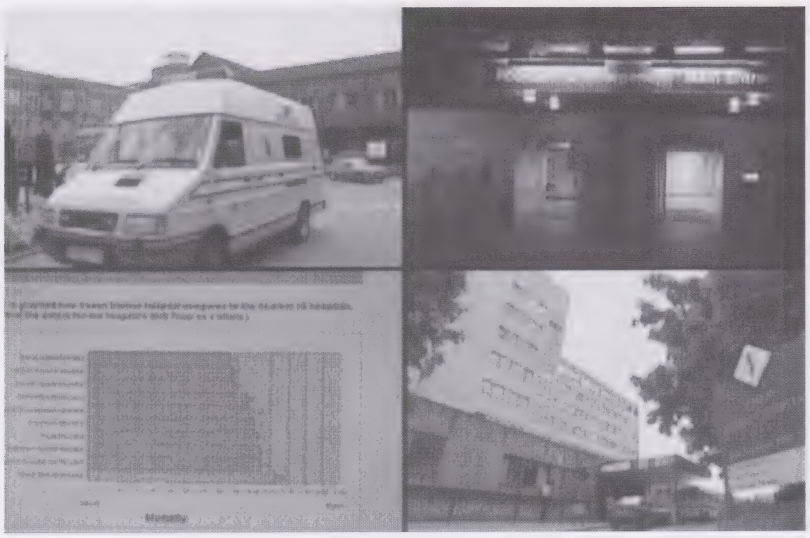

Figure 6 Can this hospital's performance be boiled down to a single number?

Statistical information of various kinds is an inescapable part of daily life. Performance indicators make up one variety of statistical information that has become increasingly prevalent and increasingly important in recent years.

Performance indicators, as the name implies, are ways of measuring the performance of an organization. A performance indicator is (usually) numerical and, in general, should relate to a stated objective of the organization involved. The broad idea is that those responsible for, or interested in, the organization should know what the organization is trying to achieve — the organization should have clear objectives — and it should be possible to measure, in a reasonably appropriate way, whether the objectives are being met.

Performance indicators in one shape or form have, of course, been around for a very long time. For instance, commercial companies have always measured their financial profits in some way or another. However, the term 'performance indicator' is relatively recent. The term itself, and the idea of wide use of measures of non-financial as well as financial performance, come from private industry, but these days most mentions of the term in the media relate to public organizations. The idea here is that performance indicators not only help public sector managers to monitor and achieve their objectives, but also allow politicians, Parliament, local authorities, and indeed any member of the public, to see how a particular organization or a particular policy is doing.

The programme explores some aspects of the use of performance indicators in the public sector in the UK. It concentrates mostly on locally-delivered services — schools, household waste collection, police, health care. The first area investigated is school 'league tables'. State schools in England have to provide information on their pupils' performance in public examinations such as GCSEs, and the resulting data are published by central government. (At the time of writing, some similar information is published for Scottish schools, but no longer for schools in Wales or Northern Ireland.) The programme concentrates mostly on one of the resulting performance indicators: the percentage of pupils, aged 15 at the start of the relevant school year, that obtain five or more GCSE passes at grades A* to C. Many parents use this information in making their choice of secondary school for their children. The programme points out that such performance indicators do not, of course, provide all the necessary information that a parent might need in making such a decision. Further, the possibility arises that a school could indulge in 'gaming' — changing its behaviour in order to improve the value of the performance indicator without actually improving its

underlying performance (though the head teacher interviewed in the programme casts doubt on how feasible this would be in practice). Also, this basic performance indicator takes no account of how able the pupils were when they entered the school. (So-called 'value added' indicators, that do take into account a measure of the pupils' initial ability, are being developed and were first published on a pilot basis for some secondary schools in 2001.)

Harvey Goldstein, though he sees uses for this sort of performance indicator in monitoring education policy, criticizes them robustly as sources of information about individual schools. He points out the problem that the published results are subject to sampling variability. That is, the actual examination results achieved in a school are not entirely determined by how good the teaching is in the school, but are also affected by chance factors. Thus, in statistical terms, all that can safely be said about a particular school is that its true underlying performance falls into a certain confidence interval. As Harvey Goldstein illustrates in a diagram, because the number of pupils in a given year group in a given school is not particularly big, the resulting confidence intervals will be quite wide. Therefore, two schools that come in quite widely separated places in the league tables may in statistical terms really be indistinguishable in terms of their actual underlying performance. (In statistical testing jargon, one could not reject the null hypothesis that their underlying performance rates were exactly the same.) In Goldstein's view, this makes the tables potentially very misleading for comparing schools. Value added tables are better in certain other respects, but still suffer from this same problem of sampling variability.

The programme then turns to performance indicators for local government. A public body known as the Audit Commission — or in full The Audit Commission for Local Authorities and the National Health Service of England and Wales — is responsible for helping local authorities and the NHS to deliver economic, efficient and effective public services. As part of this remit, it coordinates the production of a common set of performance indicators for all aspects of local government activity. These performance indicators are published annually, so that councillors, residents and local taxpayers can see how their local council is performing. Just one out of the long list of local authority performance indicators is looked at in some detail: Best Value Performance Indicator BV88, the number of collections missed per 100 000 collections of household waste. Though, like the other performance indicators, this one is tightly defined and monitored by the Audit Commission, the data are provided by the individual local authorities involved, and different authorities may not provide or collect the data on exactly the same basis. (However, it seems very unlikely indeed that, for instance, differences in data collection are responsible for the huge difference between the value of this performance indicator for the best performing and worst performing London boroughs — nearly 7000 collections missed per 100 000 in Waltham Forest and only 21 per 100 000 in Camden!)

(As a footnote to the story of this performance indicator presented in the programme, it is interesting to note that this particular performance indicator will no longer be part of the 'compulsory' list from 2002/03. The Government wrote in a consultation document, 'We consider that the indicator does not sufficiently reflect the experience and satisfaction of users with their waste collection services. Year-on-year changes in performance can be statistically insignificant and we believe that continuing collection of this data is not useful for future time series analyses, or for the target-setting purposes of local authorities themselves. Data for time series analyses on performance related to statutory targets is of more immediate value. We therefore propose to delete it and will rely on collection of the three-yearly indicator on percentage of people satisfied with waste collection.' The 'three-yearly indicator' referred to is derived from a survey of the public, who are asked among other things whether they are satisfied with the waste collection arrangements in their area. Data from such a survey were last collected in 2000/01. In London, there was practically no discernable relationship between levels of satisfaction in this survey and the performance indicator on collections missed. For example, satisfaction levels in Camden were considerably lower than those in Waltham Forest (72% very or fairly satisfied in Camden, 81% in

Waltham Forest), although the numbers of collections missed per 100 000 were not dissimilar from those given in the programme for a later year (22 in Camden, 5127 in Waltham Forest)!)

Next in the programme, Ellis Cashmore points out, from his own research, an example where the use of performance indicators in the police service led to a very undesirable unintended consequence in terms of increased racism.

David Boyle, from a think-tank called the New Economics Foundation, is a strong critic of performance indicators as they are currently used. He feels that an over-reliance on these indicators can lead to a society that is over-centralized and afraid to use common sense. However, he is certainly not against using numerical measures, and in particular the New Economics Foundation has promoted the use of performance indicators specifically decided on by the residents of local communities, to reflect and monitor their own priorities. (Boyle presents his ideas in a book, *The Tyranny of Numbers*, published in 2001.)

Finally, the programme turns to health care. Brian Jarman presents the mortality index that can be used to compare hospitals on the Dr Foster website (www.drfoster.co.uk). The data used on this website come from official sources, but the site is a private venture and uses its own methods for analysing and presenting the data. Jarman explains that it would be inappropriate simply to compare hospitals on the basis of numbers of deaths, or simple death rates, because the populations they serve are different. (Other things being equal, if one hospital is serving an older population than another, you would expect the first hospital to have higher death rates.) Also, different hospitals treat different mixes of medical conditions. The mortality index is therefore adjusted to allow for differences in age and sex in the population, for differences in the proportion of emergency admissions, and for the primary diagnosis for the patients. Brian argues that, after these adjustments, the mortality index does indeed compare like with like, so that any differences in mortality index between hospitals actually reflect the quality of care in the hospitals. (Not all health professionals or statisticians agree!) Brian Jarman also expresses the view that some of the other problems of performance indicators do not apply to the mortality index. He says that the quality of the data is high because (for instance) deaths are carefully recorded. The issue of sampling variability is not so important as in the case of schools because the number of patients involved in a hospital is typically much larger than the number of pupils in a school year. (But again, there has been some criticism of the index on both these grounds.)

In summary, performance indicators can be useful tools for evaluating policy and for making individual decisions. But they should not be used uncritically; watch out for data accuracy, for manipulation or 'gaming', for questions of sampling variability, and for unintended effects of using performance indicators.

Presenter: Jancis Robinson.

Academic consultants: Dr Paul Anand and Dr Kevin McConway.

Other contributors: Alan Davison, Professor Harvey Goldstein, Peter Wilkinson, Dr Kevin McConway, Professor Ellis Cashmore, David Boyle, Professor Brian Jarman, Carol Wells, Bill Closier, Christopher and Geraldine Balch and Dr Simon Cave.

Producer: David Berry.

Track 7 In search of certainty

Figure 7 Sir Ronald Aylmer Fisher

At the start of the twentieth century, the discipline of statistics did not exist. Just a handful of individual mathematicians and scientists were engaged in statistical work. By the beginning of the twenty-first century, statistics had grown into a major academic and practical subject which touches almost every field of science and of human endeavour.

If you were to ask any modern day statistician to name the one person who had the greatest effect on statistics in the twentieth century, the answer would almost certainly come back: R.A. Fisher! This programme explores the life and work of this great man. It also explores his legacy; while Fisher worked mostly in the first half of the twentieth century, many of the methods that Fisher invented continue to be used in modern statistical research, with applications in agriculture, medicine and industry, to name but a few. The computer revolution has not made Fisher's work obsolete. Far from it: in the programme, you see how computers make the application of Fisher's techniques — along with other important methods of statistics — happen with a click or two of a mouse.

But Fisher was not only a statistician and indeed did not call himself a statistician but a scientist. Fisher was also a great geneticist, and was hugely influential in laying the foundations of that subject too in the first half of the twentieth century. Indeed, in a throw-away remark in his book, *The Selfish Gene*, Richard Dawkins refers to Fisher as 'the greatest biologist of the twentieth century'. Not bad for one man!

The programme first explores Fisher's time (1919–1934) at Rothamsted, the world's longest running centre for agricultural research, near Harpenden in Hertfordshire. Fisher's daughter, Joan Fisher Box, takes a nostalgic journey into the past, and sees again the fields on which some of the world's longest running experiments continue to be run. Fisher was initially taken on to re-examine the mass of existing experimental data. But he soon emphasized the importance of involving the statistician right at the start of an experiment, in determining its *design*. A modern example of such a designed experiment at Rothamsted, the Winter Wheat experiment, is seen, and its analysis, though performed very quickly on a computer, still follows methods that Fisher himself invented at Rothamsted. As Michael Healy explains, Fisher's work at Rothamsted was pioneering: he had to (re-)invent the discipline of statistics, and thus laid the foundations for much of what is done today.

While Fisher's work directly for Rothamsted can be categorized as statistics, he was also given time and space to develop his own research into genetics. His work provided a mathematical framework for Darwin's evolutionary theory. In 1934, he moved to a professorship in Genetics at University College, London and in 1943, to a professorship in Genetics at Cambridge. Of course, neither of these genetics appointments put a halt to his statistics work; the two always went hand in hand.

We hear much about Ronald Aylmer Fisher's personality and character. He had a background in mathematics, which meant he liked certainty, but he also had a great interest in biological problems and realized that there is no such thing as certainty in scientific research. As Michael Healy suggests, Fisher was driven by the search for certainty in the presence of uncertainty. Fisher was, from a young age, very short-sighted. He therefore developed an extraordinary ability to perform complex mathematical and geometrical arguments in his head.

But Fisher was not a great communicator. He had a tough time getting his ideas across both to his students and to the other intellectual giants of the day. The latter often led to acrimonious public debate. An example, not in the programme, is this excerpt from Fisher's published discussion of a paper by the great Polish statistician Jerzy Neyman and colleagues in 1935, when they dared to present a paper to the Royal Statistical Society entitled 'Statistical Problems in Agricultural Experimentation': Fisher 'had hoped that Dr. Neyman's paper would be on a subject with which the author was fully acquainted, and on which he could speak with authority ... Since seeing the paper, he had come to the conclusion that Dr. Neyman had been somewhat unwise in his choice of topics'. To be fair, such acerbic remarks were fairly common in scientific debate in those days, but Fisher contributed more than his fair share!

Fisher had published his first paper as an undergraduate student at Caius College, Cambridge, in 1912, and this paper already contained the fundamentals of the method of *maximum likelihood* which is still of such huge importance today. In the programme, Alan Grafen gives an exposition of the method in general terms.

The method of maximum likelihood remains the statistical technique at the hub of the work of the team of statisticians involved with the Human Genome Project at the Sanger Centre near Cambridge, as their leader Richard Durbin explains. This project is the most important and famous modern expression of the science of genetics: the compilation of the complete genetic map for a human being. As well as providing the basis of the statistical methodology used, Anthony Edwards says that while at University College in the 1930s, Fisher was first to suggest the usefulness for disease prognosis of the development of a complete linkage map of man.

Another contemporary application of statistics in the programme is to the work of Ford engineers in developing new engine designs. It is the experimental design ideas of Fisher that are at the heart of the complex work of the development team. And, as Dean Rose explains, it is modern computers that make Fisher's methods and their extensions viable in such a context.

While Fisher's contribution to the theory of statistics ebbs and flows, to some extent, with the changing tide of ideas, there are two reasons why his reputation has been tarnished a little. One was that he entered the 1950s debate on the link between lung cancer and smoking 'on the wrong side'. The other was that he was a leading proponent of eugenics. In footage from 1937, Julian Huxley explained: 'Eugenics seeks to apply the known laws of heredity so as to prevent the degeneration of the race and improve its inborn qualities'.

Nonetheless, Fisher's influence on statistics — and hence on knowledge and research in almost every discipline — remains pervasive, as does his influence on evolutionary biology. 'Fisher is the greatest of them all, he's the number one. You can't really think of anybody who's more important in the history of statistics' (Stephen Senn). 'He was a genius, of course' (Joan Fisher Box).

Academic consultant: Professor Chris Jones.

Contributors: Joan Fisher Box, Professor Michael Healy, Professor Anthony Edwards, Professor Stephen Senn, Dr Alan Grafen, Dr Richard Durbin and Dean Rose.

Narrator: Glenda Jackson.

Producer: Liz Gray.

The paper used in this publication contains pulp sourced from forests independently certified to the Forest Stewardship Council®(FSC®) principles and criteria. Chain of custody certification allows the pulp from these forests to be tracked to the end use (see www.fsc-uk.org).

Printed in the United Kingdom by Hobbs the Printers Ltd, Totton, Hampshire